THE MATHEMATICAL REPRESENTATION OF FREQUENCY DISTRIBUTIONS

By Harry C. Carver, *University of Michigan*

Section III

DIFFERENCE EQUATION GRADUATION

Certain geometrical properties of unimodal frequency distributions suggest that any associated frequency function may be represented as a solution of the difference equation.

$$\frac{\Delta y_x}{\Delta_x} = \frac{y_x(a-x)}{f(x)} \tag{1}$$

since

(a) if there be one mode only there must be a value of $x=a$ for which $\Delta y_x = 0$, and

(b) towards the extremes, the finite difference between two successive ordinates must approach zero as y_x diminishes in value.

The balance of the difference equation of the unknown theoretical law of distribution may be represented by a function, $f(x)$, appearing in the denominator.

We shall now *assume* that $f(x)$ may be expanded in a power series which in practice is found to be rapidly convergent. The merits of this important assumption will be discussed briefly later.

Expressing (1) then as

$$(b_0+b_1x+b_2x^2+ \ . \ . \ . \ )\Delta y_x = (a-x)y_x \cdot \Delta x,$$

multiplying through by x^n and summing with respect to x yields

$$b_0\Sigma x^n\Delta y_x + b_1\Sigma x^{n+1}\Delta y_x + b_2\Sigma x^{n+2}\Delta y_x + \ . \ . \ . \ = (a\Sigma x^n y_x - \Sigma x^{n+1}y_x)\Delta x. \tag{2}$$

* Sections I and II of this paper appeared in the June issue of the Quarterly Publications.

If the range of the distribution be from $x=-\infty$ to $x=\infty$, we have by finite integration by parts

$$\sum_{x=-\infty}^{\infty} x^n \Delta y_x = x^n y_x - \Sigma \left\{ (x+\Delta x)^n - x^n \right\} y_{x+\Delta x} \Big]_{x=-\infty}^{\infty}$$

$$= -\Sigma \left\{ x^n - (x-\Delta x)^n \right\} y_x \Big]_{x=-\infty}^{\infty}$$

$$= -nC_1 \Sigma x^{n-1} y_x \Delta x + {}_nC_2 \Sigma x^{n-2} y_x (\Delta x)^2 - \ \ldots \tag{3}$$

When dealing with distributions of graduated variates, Δx should be permitted to approach zero as a limit.

Thus, equations (1), (2), and (3) become

$$\frac{1}{y}\frac{dy}{dx} = \frac{a-x}{b_0+b_1x+b_2x^2+\ \ldots} \tag{1a}$$

$$b_0\int x^n dy + b_1\int x^{n+1}dy + b_2\int x^{n+2}dy + \ \ldots = a\int x^n y dx - \int x^{n+1} y dx \tag{2a}$$

$$\int x^n dy = -N \cdot n\nu'_{n-1}. \tag{3a}$$

If we choose the mean of the distribution as origin and give n in (2a) successively the values 0, 1, 2, . . . we obtain

$$\left\{\begin{array}{rrrrl} a & +b_1 & + & \ldots & =0 \\ & b_0 & +3\nu_2 b_2 + & \ldots & =\nu_2 \\ \nu_2 a & +3\nu_2 b_1 + 4\nu_3 b_2 + & & \ldots & =\nu_3 \\ \nu_3 a + 3\nu_2 b_0 + 4\nu_3 b_1 + 5\nu_4 b_2 + & & & \ldots & =\nu_4 \end{array}\right. \tag{4a}$$

etc.

For distributions of graduated variates we take the common difference between any two successive class magnitudes as the unit for x, (thus $\Delta x=1$) and giving n successively the values 0, 1, 2, . . . as before we obtain from (2) and (3)

$$\left\{\begin{array}{rrrrl} a & & +b_1 & -b_2+ & \ldots =0 \\ & b_0 & -b_1+ & (3\nu_2+1)b_2+ & \ldots =\nu_2 \\ \nu_2 a & -b_0+ & (3\nu_2+1)b_1+ & (4\nu_3-6\nu_2-1)b_2+ & \ldots =\nu_3 \\ \nu_3 a + (3\nu_2+1)b_0 + & (4\nu_3-6\nu_2-1)b_1 + & (5\nu_4-10\nu_3+10\nu_2+1b_2+ & & \ldots =\nu_4 \end{array}\right. \tag{4}$$

etc.

It should be noted that the moments of (4a) are defined by

$$\nu_n = \int x^n y dx$$

whereas those of (4) are given by

$$\nu_n = \Sigma x^n y_x.$$

A simultaneous solution of equations (4*a*) determines the constants of the *differential* equation (1*a*), which on integration produce Pearson's system of Generalized Probability Curves.

Equations (4) likewise determine the constants for the corresponding *difference* equation (1), when $\Delta x = 1$.

If for a particular distribution the series $b_0 + b_1 x + b_2 x^2 + \ldots$ converges rapidly and it is possible to neglect all terms containing powers of x greater than the second, solutions of (4*a*) and (4) yield, letting

$$\beta_1 = \frac{\nu_3^2}{\nu_2^3} \text{ and } \beta_2 = \frac{\nu_4}{\nu_2^2}.$$

TABLE I

Differential values	Const.	Difference values
$\dfrac{-\frac{\nu_3}{\nu_2}(\beta_2+3)}{2(5\beta_2-6\beta_1-9)}$	a	$\dfrac{-\frac{\nu_3}{\nu_2}\left(\beta_2+3-\frac{1}{\nu_2}\right)}{2\left(5\beta_2-6\beta_1-9+\frac{1}{\nu_2}\right)}-\frac{1}{2}$
$\dfrac{\nu_2(4\beta_2-3\beta_1)}{2(5\beta_2-6\beta_1-9)}$	b_0	$\dfrac{\nu_2\left(4\beta_2-3\beta_1-\frac{1}{\nu_2}\right)}{2\left(5\beta_2-6\beta_1-9+\frac{1}{\nu_2}\right)}-a$
$-a$	b_1	b_2-a
$\dfrac{2\beta_2-3\beta_1-6}{2(5\beta_2-6\beta_1-9)}$	b_2	$\dfrac{2\beta_2-3\beta_1-6+\frac{1}{\nu_2}}{2\left(5\beta_2-6\beta_1-9+\frac{1}{\nu_2}\right)}$

If the series, $f(x)$, converges so rapidly that the term $b_2 x^2$ may also be neglected, that is in cases where the value of b_2 is not appreciably greater than its probable error, we have

TABLE II

Differential values	Const.	Difference values
$-\dfrac{\nu_3}{2\nu_2}$	a	$\dfrac{-\nu_3}{2\nu_2}-\dfrac{1}{2}$
ν_2	b_0	ν_2-a
$-a$	b_1	$-a$

In application, the difference equation has certain advantages over the differential equation. Thus, a knowledge of the values of the constants of the differential equation permits us to compute the theoretical ordinates only after the integration of the differential equation, the constant of integration being determined by imposing the condition that the sums of the graduated and ungraduated frequencies must be equal.

The difference equation, however, requires no integration, since a knowledge of the constants permits us to compute first all necessary values of $\frac{\Delta y_x}{y_x}$ and then $\frac{y_{x+1}}{y_x}$ from which ordinates proportional to those required may be computed by successive multiplication. The condition that the sum of the graduated frequencies must equal that of the ungraduated determines the proper proportional factor.

For numerical illustrations of the use of the difference equation method for graduating complete distributions as well as "stumps" of distributions, reference may be made to "On the Graduation of Frequency Distributions" by the writer.*

Section IV

Application of the Hypergeometric Series

$$\text{(A)} \quad u_x = {}_{pn}C_{r-x} \quad {}_{qn}C_x$$

$$\text{(B)} \quad u_x = {}_{pn}H_{r-x} \quad {}_{qn}H_x$$

Critical investigations of the variations which are found to exist in apparently homogeneous statistical data have led to the development of various theories of frequency distribution.

One of the first of these, known to biologists as Quetelet's Law, states that the distribution of individuals ranked according to some common character in a frequency series may be represented by the successive terms of the expansion of the point binomial

$$N(p+q)^r,$$

that is to say

$$\text{(1)} \qquad N\left\{p^r + {}_rC_1 p^{r-1}q + {}_rC_2 p^{r-2}q^2 + \ \ldots\ + q^r\right\}$$

where $p+q=1$

N = the total frequency of the distribution.

r = an integer, representing, therefore, one less than the number of classes in the theoretical distribution.

* Published in the *Proceedings of the Casualty Actuarial and Statistical Society of America*, vol. vi, part 1, No. 13.

A student of probabilities, however, is more apt to associate the name of Bernoulli with this series, since the successive terms are merely the frequency expectations of $r, r-1, r-2, \ldots 0$ occurances in N trials of r independent events each, where the probability of the happening of each event is designated by p and of its non-happening by q.

The difference equation of Quetelet's Law, taking the position of the first term as origin, is

$$\frac{y_{x+1}}{y_x} = \frac{N \cdot {}_rC_{x+1} p^{r-x-1} q^{x+1}}{N \cdot {}_rC_x p^{r-x} q^x} = \frac{(r-x)q}{(x+1)p}$$

or

$$\frac{\Delta y_x}{y_x} = \frac{rq-p-x}{(x+1)p}.$$

If the origin be now shifted to the mean, rq units distant, the new difference equation becomes

$$\frac{\Delta y_x}{y_x} = \frac{-p-x}{p(1+rq)+px} \tag{2}$$

which is of the form

$$\frac{\Delta y_x}{y_x} = \frac{a-x}{b_0+b_1 x}.$$

From the values of the constants of the difference equations given in Table II we have

$$p = \frac{\nu_2+\nu_3}{2\nu_2}, \quad q = \frac{\nu_2-\nu_3}{2\nu_2}, \quad r = \frac{4\nu_2^3}{\nu_2^2-\nu_3^2}. \tag{3}$$

From the above we see that if p, q, and r are to have any real significance, the absolute value of ν_3 must be less than ν_2. Otherwise r would be negative and either p or q would be greater than unity.

An important limit of Quetelet's Law is obtained by permitting q to approach zero and r infinity in such a manner that the product rq, representing the distance from the origin of the series to the mean, remains a constant and equal to m.

This limit,

$$\frac{N}{e^m}\left[1+m+\frac{m^2}{\underline{|2}}+\ldots+\frac{m^x}{\underline{|x}}+\ldots\right] \tag{4}$$

is known as Poisson's Exponential Binomial Limit, and is often referred to as the Law of Small Numbers.

The criterion for this series is obviously $\nu_2 = \nu_3$.

In the *Philosophical Transactions of the Royal Society of London* (vol. 186, part 1, p. 360), Pearson presents a generalized series which is more general and powerful than the point binomial mentioned above.

It may be developed as follows:

If from a bag containing pn black and qn white balls, r balls are withdrawn without replacements, the chances that the r balls withdrawn will contain $r, r-1, r-2, \ldots 2, 1, 0$ black balls are given by the successive terms of the hypergeometric series

$$(5) \qquad \frac{1}{{}_nC_r}\left\{{}_{pn}C_r + {}_{pn}C_{r-1}\,{}_{qn}C_1 + {}_{pn}C_{r-2}\,{}_{qn}C_2 + \ldots + {}_{qn}C_r\right\}.$$

The difference equation of this series, referred to the mean which is rq units distant from the first term, is

$$(6) \qquad \frac{\Delta y_x}{y_x} = \frac{(r\,\overline{2p-1} - pn - 1) - (n+2)x}{(r\,\overline{1-p}+1+x)(p\,\overline{n-r}+1+x)}.$$

Comparing equation (6) with (4) of Section III, we obtain the following for the hypergeometric series $u_x = {}_{pn}C_{r-x}\,{}_{qn}C_x$:

$$(7) \quad \begin{cases} \nu_2 = rpq\dfrac{n-r}{n-1} \\ \nu_3 = \nu_2(p-q)\dfrac{n-2r}{n-2} \\ \nu_4 = \dfrac{\nu_2}{(n-2)(n-3)}\left\{3(n-1)(n+6-\dfrac{2}{pq})\nu_2 + n(n+1) - 6pqn^2\right\}. \end{cases}$$

But here again we find that for distributions of integral variates our results are unintelligible unless ν_3 is in absolute value less than ν_2.

Again, since for this series $b_2 = \dfrac{1}{n+2}$ we have from Table I,

$$(8) \qquad n = \frac{6(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6 + \dfrac{1}{\nu_2}};$$

and a few trials will convince one that for many of the distributions that are met in practice this solution yields a negative value for n, and that this occurs when $\nu_3 > \nu_2$.

If we now consider the hypergeometric series

$$u_x = {}_{pn}H_{r-x} \cdot {}_{qn}H_x$$

where ${}_nH_r$ denotes the number of combinations of n things taken r at a time when repetitions are allowed, *i. e.*, ${}_nH_r = {}_{n+r-1}C_r$, we have

$$\frac{\Delta y_x}{y_x} = \frac{(r\overline{1-2p}+1-pn)-(n-2)x}{(r\overline{1-p}+1+x)(p\overline{n+r}-1-x)}, \tag{9}$$

where the mean, which is also at rq, is taken as the origin.

Comparing equation (9) with (4) of Section III we obtain for the series $u_x = {}_{pn}H_{r-x} \cdot {}_{qn}H_x$

$$\left\{\begin{array}{l} \nu_2 = rpq\dfrac{n+r}{n+1} \\ \nu_3 = \nu_2(p-q)\dfrac{n+2r}{n+2} \\ \nu_4 = \dfrac{\nu_2}{(n+2)(n+3)}\left\{3(n+1)(n-6+\dfrac{2}{pq})\nu_2 + n(n-1) - 6pqn^2\right\}. \end{array}\right. \tag{10}$$

It may now be noted that equations (9) and (10) may be obtained directly from (6) and (7) by replacing n in the latter by $(-n)$.

If for convenience we designate the point binomial or Bernoulli's series by Series B, the hypergeometric series $u_x = {}_{pn}C_{r-x\,qn}C_x$ as Series C, and $u_x = {}_{pn}H_{r-x\,qn}H_x$ as Series H, we see that Series H may be used for those distributions for which Series B and C are meaningless.

An analysis of the means and dispersions of these three series is enlightening. From the following table we note that although their

TABLE III

Series	Mean	Dispersion
C	rq	$rpq\frac{n-r}{n-1}$
B	rq	rpq
H	rq	$rpq\frac{n+r}{n+1}$

means are identical the dispersion of Series C is always less and that of Series H greater than the Bernoullian dispersion. Moreover, as n approaches infinity, the dispersions of Series C and H approach the Bernoullian as a limit from opposite sides.

Inasmuch as the Bernoulli Series, because of its rather extensive degree of freedom, is itself a powerful "closed" graduation function, it follows that the combination of these three series—affording an addi-

tional continuous degree of freedom—is capable of graduating practically any unimodal distribution.

These considerations throw an interesting light on the convergence of $f(x) = b_0 + b_1x + b_2x^2 + \ . \ . \ .$ in the denominator of either the difference or differential equation. If we stop with b_1x the freedom is restricted to that of a point binomial, and the addition of b_2x^2 increases the freedom to at least that of the hypergeometric series.

At the present time tables, based on formulae (7) and (10), are being prepared which will enable one to obtain by inspection the proper values of p, q, r, and n when the values of the moments are known. By this method it is hoped that a simple method of graduating frequency distributions may be available, and, what is more important, that something may be accomplished in the direction of classifying distributions.